Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Letian Wang March 20, 2017

A Study of Benford's Law for the Values of Arithmetic Functions

By

Letian Wang

Ken Ono, Ph.D.

Advisor

Department of Mathematics and Computer Science

Ken Ono, Ph.D.

Advisor

John F. R. Duncan, Ph.D.

Committee Member

Jed Brody, Ph.D.

Committee Member

2017

A Study of Benford's Law for the Values of Arithmetic Functions

By

Letian Wang

Ken Ono, Ph.D.

Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Mathematics and Computer Science

2017

Abstract

A Study of Benford's Law for the Values of Arithmetic Functions
By Letian Wang

Benford's Law characterizes the distribution of initial digits in large datasets across disciplines. Since its discovery by Simon Newcomb in 1881, Benford's Law has triggered tremendous studies. In this paper, we will start by introducing the history of Benford's Law and discussing in detail the explanations proposed by mathematicians on why various datasets are Benford. Such explanations include the Spread Hypothesis, the Geometric, the Scale-Invariance, and the Central Limit explanations.

To rigorously define Benford's Law and to motivate criteria for Benford sequences, we will provide fundamental theorems in uniform distribution modulo 1. We will state and prove criteria for checking uniform distribution, including Weyl's Criterion, Van der Corput's Difference Theorem, as well as their corollaries.

We will then introduce the logarithm map, which allows us to reformulate Benford's Law with uniform distribution modulo 1 studied earlier. We will start by examining the case of base 10 only and then generalize to arbitrary bases.

Finally, we will elaborate on the idea of *good* functions. We will prove that good functions are Benford, which in turn enables us to find a new class of Benford sequences. We will use this theorem to show that the partition function $p(n)$ and the factorial sequence $n!$ follow Benford's Law.

A Study of Benford's Law for the Values of Arithmetic Functions

By

Letian Wang

Ken Ono, Ph.D.
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Mathematics and Computer Science

2017

Acknowledgment

I would like to thank Dr. Ken Ono
for his expertise, guidance, and tremendous support throughout my project;

I am grateful for Dr. John Duncan and Dr. Jed Brody
for their inspiration and encouragement.

Table of Contents

1 Introduction **1**

1.1 A Brief History 2

1.2 Examples and Explanations 4

1.2.1 The Spread Hypothesis 7

1.2.2 The Geometric Explanation 8

1.2.3 The Scale-Invariance Explanation 9

1.2.4 The Central Limit Explanation 9

1.3 What's Next 11

2 Uniform Distribution **12**

2.1 Uniform Distribution Modulo 1 12

2.2 Weyl's Criterion 16

2.3 Difference Theorem 20

3 Mathematical Framework for Benford's Law **24**

3.1 Benford's Law in Base 10 24

3.2 Generalization to All Bases 26

3.3 Asymptotic Property 29

4 Benford Arithmetic Functions **30**

4.1 Good Functions 30

4.2 The Partition Function $p(n)$ is Benford 31

4.3 Factorials are Benford 33

Bibliography **34**

Chapter 1 *Introduction*

Benford's Law is a statistical characterization of initial digits distribution over datasets. Originated from the observation that the beginning pages of a logarithm table wear out faster than the later ones, Benford's Law finds broad applications in fraud detection and image processing, generating hundreds of research papers in Accounting, Computer Science, Engineering, Statistics, and many other disciplines [7]. In this chapter, we will begin by examining the historical development of Benford's Law. We will then give examples of datasets where the law is satisfied or unsatisfied anecdotally and provide explanations for both cases, thereby motivate the result of Anderson, Rolen and Stoehr on partition functions [1].

Before we embark, a few definitions are necessary.

Definition 1. (*Significand and Mantissa*) We know that any real number x can be written in scientific notation as $S(x) \cdot 10^k$, where $S(x) \in [1, 10)$. We define $S(x)$ as the *significand* and k the *exponent* of x, both in base 10. We will call the integer part of $S(x)$, or $\lfloor S(x) \rfloor$, the *leading digit* of x. We will call the fractional part of $\log_{10} x$, $\mathrm{frac}\big((\log_{10} x)\big)$ or $(\log_{10} x \bmod 1)$, the *mantissa* of x in base 10.

For example, suppose $x = 1234.56$, then $S(x) = 1.23456$, $k = 3$ by scientific notation, and the leading digit of x is $\lfloor S(x) \rfloor = 1$. The mantissa of x in base 10 is therefore $\log_{10} 1234.56 \bmod 1 = 3.0915\cdots \bmod 1 = 0.0915\cdots$. Notice that all mantissas fall in the interval $[0, 1)$, and every real number within the same interval is the mantissa of infinitely many real numbers. In this chapter, we will only consider cases in base 10. In Section 3.2, we will generalize our investigation to arbitrary bases.

1.1 A Brief History

With the aforementioned definitions in mind, Benford's Law describes the biased distribution of the leading digits of numbers in datasets. Should the leading digits be uniformly distributed, the probability of numbers starting with 1 through 9 should be identically $11.\dot{1}\%$. In reality, however, the distribution of leading digits is largely biased.

The first observer of this phenomenon is the Canadian mathematician Simon Newcomb. In his 1881 article *Note on the Frequency of Use of the Different Digits in Natural Numbers*, Newcomb observed that the opening pages of logarithm tables wear out much faster than do the later pages. He claims that "the [leading digit] is oftener 1 than any other digit, and the frequency diminishes up to 9"[8]. Newcomb quantified the bias in Table 1, where D_1 represents the leading digits 1 through 9, and $P(D_1)$ denotes the probability that the leading digit is D_1.

D_1	1	2	3	4	5	6	7	8	9
$P(D_1)$	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0458

Table 1. *Newcomb's quantification for the distribution of leading digits in nature*

Newcomb went on to conjecture that "the law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable". We will explain Newcomb's claim in detail in Section 3.1 using logarithmic mapping and uniform distribution.

Newcomb's observation, however, did not receive immediate attention from the academia. It was the study of the American engineer and physicist Frank Benford published 57 years later that triggered thorough research on this topic. In his 1938 paper *The Law of Anomalous Numbers*, Benford explored 20 drastically different datasets ranging from population to atomic weights and counted the occurrence of each leading digit [2]. Benford's result, reproduced in Table 2 on page 4, further confirmed that $P(D_1)$ is not uniformly distributed among 1 to 9. It is apparent from the table that $P(1)$ tends to be the largest while $P(9)$ tends to be the smallest. In addition, even though the biased distributions coincide with Newcomb's conjecture for many datasets, the two differ significantly in other cases. By taking the average amongst all 20 datasets, the amalgamated probabilities listed in the bottom row of Table 2 are closer to those listed in Table 1.

In view of the above observations, we now sketch a mathematical definition of Benford's Law, named in honor of Frank Benford.

Definition 2. (*Benford's Law*) A sequence $\{x_n\}$ satisfies Benford's Law if the probabilistic condition $P(D_1) = \log_{10}(D_1 + 1) - \log_{10}(D_1)$ holds in $\{x_n\}$, in which case we say that $\{x_n\}$ is Benford.

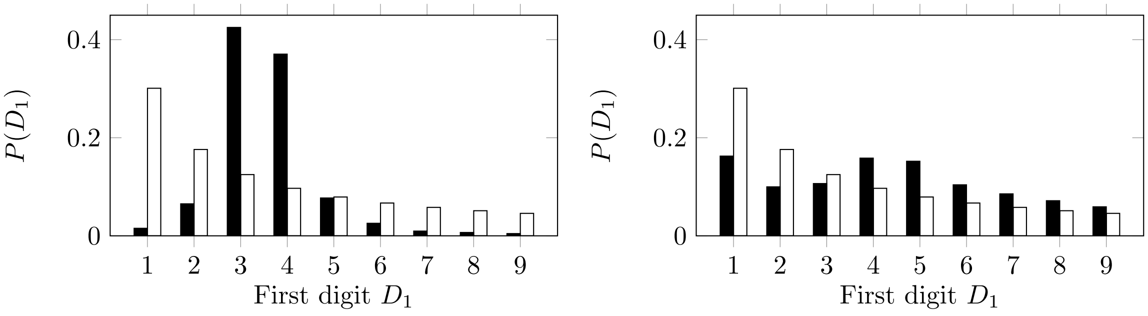
Dataset	$P(1)$	$P(2)$	$P(3)$	$P(4)$	$P(5)$	$P(6)$	$P(7)$	$P(8)$	$P(9)$	*Size*
Rivers, Area	0.310	0.164	0.107	0.113	0.072	0.086	0.055	0.042	0.051	335
Population	0.339	0.204	0.142	0.081	0.072	0.062	0.041	0.037	0.022	3,259
Constants	0.413	0.144	0.048	0.086	0.106	0.058	0.010	0.029	0.106	104
Newspaper	0.300	0.180	0.120	0.100	0.080	0.060	0.060	0.50	0.050	100
Specific Heat	0.240	0.184	0.162	0.146	0.106	0.041	0.032	0.048	0.041	1,389
Pressure	0.296	0.183	0.128	0.098	0.083	0.064	0.057	0.044	0.047	703
H.P. Lost	0.300	0.184	0.119	0.108	0.081	0.070	0.051	0.051	0.036	690
Mol. Wgt.	0.267	0.252	0.154	0.108	0.067	0.051	0.041	0.028	0.032	1,800
Drainage	0.271	0.239	0.138	0.126	0.082	0.050	0.050	0.025	0.019	159
Atomic Wgt.	0.472	0.187	0.055	0.044	0.066	0.044	0.033	0.044	0.055	91
$n^{-1}, \sqrt{n}$	0.257	0.203	0.097	0.068	0.066	0.068	0.072	0.080	0.089	5,000
Design	0.268	0.148	0.143	0.075	0.083	0.084	0.070	0.073	0.056	560
Digest	0.334	0.185	0.124	0.075	0.071	0.065	0.055	0.049	0.042	308
Cost Data	0.324	0.188	0.101	0.101	0.098	0.055	0.047	0.055	0.031	741
X-Ray Volts	0.279	0.175	0.144	0.090	0.081	0.074	0.051	0.058	0.048	707
Am. League	0.327	0.176	0.126	0.098	0.074	0.064	0.049	0.056	0.030	1,458
Black Body	0.310	0.173	0.141	0.087	0.066	0.070	0.052	0.047	0.054	1,165
Address	0.289	0.192	0.126	0.088	0.085	0064	0.056	0.050	0.050	342
$n, n^2, \cdots, n!$	0.253	0.160	0.120	0.100	0.085	0.088	0.068	0.071	0.055	900
Death Rate	0.270	0.186	0.157	0.094	0.067	0.065	0.072	0.048	0.041	418
Average	0.306	0.185	0.124	0.094	0.080	0.064	0.051	0.049	0.047	1,011

Table 2. *Benford's result on 20 datasets* (*Table 1.2 in* [7])

The definition is a direct characterization of Newcomb's conjecture. Note that since $P(D_1)$ in Definition 2 is irrational and that $P(D_1)$ for finite sequences is always rational, we assume that $\{x_n\}$ is infinite in size. A generalization of this definition will be discussed after Definition 4 in Section 3.2.

1.2 Examples and Explanations

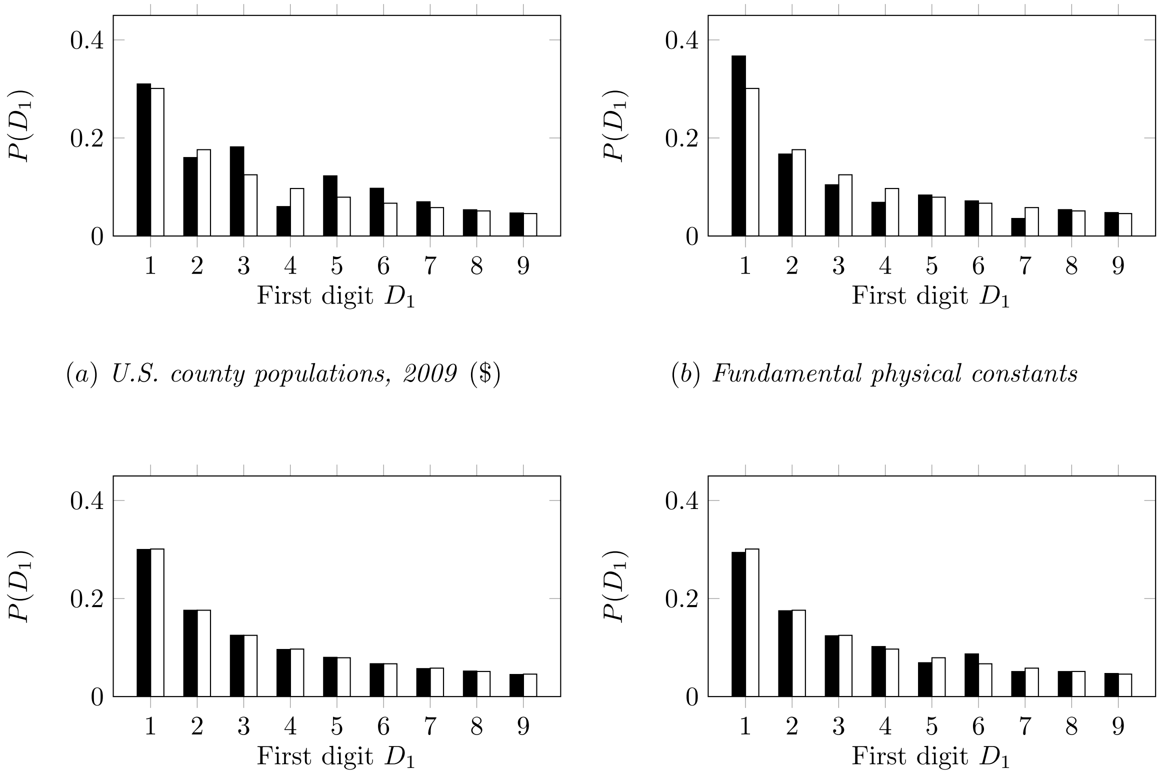
Before delving into the reasons why many datasets follow Benford's Law, we begin by exhibiting 2 datasets that do not. Plotted in Figure 1 (a) is the daily NYSE group dollar volume ($) over the 1,762 trading days from 2010 to 2016 (in black) versus Benford's Law (in white). This dataset clearly differs from Benford's Law in that d is heavily concentrated around 3 and 4 while $P(1)$ is infinitesimal. A reason for such a distribution is that transaction volumes tend to fluctuate by only a small magnitude, leaving the first digits clustered in a tiny interval [7]. Figure 1 (b) shows the distribution of land areas of 3,146 U.S. counties (mile^2) published by the U.S. Census Bureau. One possible explanation to such distribution is that early counties on the east coast are small and easy to manage by the colonial government. As the country extend to the west, county sizes become increasingly larger. Interestingly, changing the units of Figure 1 (b) from mile^2 to km^2 and acre results in different distributions, although none of which follow Benford's Law. We will revisit this observation in Subsection 1.2.3.



(*a*) *NYSE group dollar volume, 2010 ∼ 2016* ($) (*b*) *U.S. county land areas, 2010* (mile^2)

Figure 1. *Examples of datasets* (*black*) *that do not obey Benford's Law* (*white*)

We now look at examples of datasets that seem to follow Benford's Law. Plotted in Figure 2 on page 5 are first-digit distributions of (a) U.S. population by counties in 2009, (b) 335 fundamental physical constants listed on the website of National Institute of Standards and Technology (NIST), (c) first 1000 Fibonacci numbers, and (d) first 1000 factorials starting with 1!. It is visually clear to us that all 4 datasets are approximately Benford. We will refer to these examples in the subsections to come.



(*a*) *U.S. county populations, 2009* ($)

(*b*) *Fundamental physical constants*

(*c*) *The first 1000 Fibonacci numbers*

(*d*) *The first 1000 factorials*

Figure 2. *Examples of datasets* (*black*) *that appear to obey Benford's Law* (*white*)

Before we survey various explanations for why so many datasets appear to follow Benford's Law, we state a working criterion for being Benford. A formal statement of Proposition 1.1 will be introduced later in this paper after defining the logarithm map. We will also delay the proof to Section 3.1.

Proposition 1.1. (*A Working Criterion for Benford*) A sequence $\{x_n\}$ is Benford if and only if $\{(\log_{10} x_n) \bmod 1\}$ is uniformly distributed over $[0, 1)$.

1.2.1 The Spread Hypothesis

The Spread Hypothesis argues that a sequence is likely to be Benford if it spans over multiple orders of magnitude, as in the cases of (b), (c), and (d) of Figure 2. Although counterexamples such as $\{10, 100, \cdots, 10^{1000}\}$ can be easily constructed, the hypothesis serves as a convenient rule of thumb. An explanation of the Spread Hypothesis is given by William Feller [3, 7]. The strategy is to prove that Proposition 1.1 is likely true for spread out sequences. We sketch his arguments as follows.

Suppose that the random variable X is continuous on $\mathbb{R}^+$ with probability density function $f_X(x)$. Suppose also that the anti-derivative of $f_X(x)$ is $F_X(x)$. Then the cumulative distribution function of X is $P(X < x) = \int_0^x f_X(x)\,dx = F_X(x)$. Let $Y = \log_{10} X$, then cumulative distribution function of Y is $P(Y < y) = P(\log_{10} X < y) = P(X < 10^y) = F_X(10^y)$. Differentiating both sides, we get the probability density function for Y as $f_Y(y) = f_X(10^y)10^y \ln y$. The probability that $(Y \bmod 1) \in [a, b)$ is simply the summation of $P\big((Y \bmod 1) \in [a+n, b+n)\big)$ over $n \in \mathbb{N}$, or

$$P\big((Y \bmod 1) \in [a,b)\big) = \sum_{n=-\infty}^{\infty} \int_{a+n}^{b+n} f_Y(y)\,dy = \sum_{n=-\infty}^{\infty} \int_{a+n}^{b+n} f_X(10^y)10^y \ln y\,dy.$$

By Fubini's Theorem, we can switch the orders of integration and summation so that

$$P\big((Y \bmod 1) \in [a,b)\big) = \int_a^b \sum_{n=-\infty}^{\infty} f_X(10^{y+n})10^{y+n} \ln y\,dy. \tag{1.1}$$

Feller shows that if variable X spreads over multiple orders of magnitude, then the integrand $\sum_{n=-\infty}^{\infty} f_X(10^{y+n})10^{y+n} \ln y$ in (1.1) is close to uniform over $[0, 1)$.

1.2.2 The Geometric Explanation

In his paper [2], Benford provided an explanation of Benford's Law based on geometric series. We illustrate his idea by looking at the population growth of an imaginary county at a fixed rate of $r = 2\%$ per year. Intuitively, it would take longer for the population to increase from 1 million to 2 million than from 9 million to 10 million. A quantification listed in Table 3 suggests that population sizes taken at equal time intervals conform with Benford's Law (D_1 is the leading digit of population size). For instance, it would take a total of 116.2767 years for the population to increase from 1 million to 10 million, during which the first 35.0029 years are spent to grow the population from 1 million to 2 million. The percentage of time when population is in the interval $[10^6, 2 \cdot 10^6)$ is precisely 30.10%, as in Benford's Law.

D_1	1	2	3	4	5	6	7	8	9
Years Spent	35.0029	20.4754	14.5275	11.2684	9.2069	7.7844	6.7431	5.9478	5.3205
$P(D_1)$	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0458

Table 3. *Population growth at a fixed rate of* 2% *per year.*

More generally, for any fixed growth rate r,

$$\begin{aligned} P(D_1) &= \frac{\text{Time to grow from } D_1 \cdot 10^6 \text{ to } (D_1+1) \cdot 10^6}{\text{Time to grow from } 10^6 \text{ to } 10^7} \\ &= \frac{\log_{1+r} \frac{D_1+1}{D_1}}{\log_{1+r}(10)} \\ &= \log_{10}(D_1+1) - \log_{10}(D_1), \end{aligned}$$

satisfying Definition 2. This explains why Fibonacci numbers in Figure 2 (c) follow Benford's Law, since the n-th Fibonacci number $F_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n \approx \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n$, approaching a geometric series for large n.

1.2.3 The Scale-Invariance Explanation

As mentioned earlier in Figure 1 (b) that changing the units of the U.S. county land areas result in different first-digit distributions, but none of them is Benford. On the other hand, scaling the first 1000 factorials by a factor of 10 would not change its first-digit distribution and the augmented sequence is still Benford. This observation leads to another reason for the prevalence of Benford's Law because many datasets are scale-invariant. More precisely, if the first digits of the random variables X and $Y = cX$ are the same for every real constant c, then X must follow Benford's Law.

A brief explanation is as follows. Assume the conditions on X and Y. Proposition 1.1 reduces our task to showing that $Z = (\log_{10} X) \bmod 1$ is uniform. We let $X' = \log_{10} X$ and $Y' = \log_{10} Y = \log_{10} c + \log_{10} X = \log_{10} c + X'$. We claim that the probability density function of $Z = X' \bmod 1$, denoted by $f_Z(z)$, must be constant and proceed to prove by contradiction. Should $f_Z(z)$ be non-constant, it must achieve its maxima and minima at some values $f_Z(z)_{max} = \log a$ and $f_Z(z)_{min} = \log b$, where $a, b \in [1, 10)$ and $a > b$. Let $c = \dfrac{a}{b}$, then X' is most likely to be congruent to $\log_{10} a \pmod 1$ whereas Y' is least likely to be congruent to $\log_{10} c + \log_{10} b \equiv \log_{10} a \pmod 1$. The result contradicts the assumption that X' and Y' share the same first-digit distribution. Hence, $f_Z(z)_{max}$ must be constant and X must be Benford.

1.2.4 The Central Limit Explanation

Another explanation of Benford's Law utilizes the Central Limit Theorem repeated in Theorem 1.2. This explanation addresses the fact that many quantities in the real world are products of other quantities, such as the weight a cuboid ($\rho xyzg$, where ρ is the density, x, y, z are dimensions, and $g = 9.8$ ms^{-2} is the gravitational acceleration).

Theorem 1.2. (*Central Limit Theorem, Theorem 4.3.2 in* [6]) Let $W_1, W_2, \cdots$ be an infinite sequence of independent random variables, each with the same distribution. Suppose that the mean μ and the variance σ^2 of $f_W(w)$ are both finite. For any numbers a and b,

$$\lim_{n\to\infty} P\left(a \leqslant \frac{W_1 + \cdots + W_n - n\mu}{\sqrt{n}\sigma} \leqslant b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2}\, dz.$$

More precisely, suppose that

$$X = X_1 X_2 \cdots X_n.$$

Taking the logarithm of both sides and letting it equal to Z, one obtains that

$$Z = \log_{10} X = \log_{10} X_1 + \log_{10} X_2 + \cdots + \log_{10} X_n.$$

If X_i's are identically distributed independent random variables, then so are $\log_{10} X_i$. Let $Z_i = \log_{10} X_i$, and suppose that it has mean μ and variance σ^2. Then, by Theorem 1.2, Z tends to normal distribution with mean $n\mu$ and variance $n\sigma^2$. Let k be an integer, then

$$\begin{aligned} P(D_1) &= P\big(\log_{10}(D_1+1) \leqslant Z \bmod 1 \leqslant \log_{10} D_1\big) \\ &= \lim_{n\to\infty} \int_{\log_{10} D_1 + k}^{\log_{10}(D_1+1)+k} e^{-\frac{(z-n\mu)^2}{2(n\sigma)^2}} dz \\ &= \log_{10}(D_1+1) - \log_{10} D_1 \end{aligned}$$

Hence, by Definition 2, $X = 10^Z$ tends to be Benford. Moreover, by our discussion, X becomes more and more Benford as n gets larger.

1.3 What's Next

We conclude this chapter with a grand view of the following chapters.

Chapter 2 will provide the preliminary background of uniform distribution modulo 1. We will state and prove Weyl's Criterion, Van der Corput's Difference Theorem, as well as their respective corollaries.

In Chapter 3, we will establish a rigorous reformulation of Benford's Law based on uniform distribution. We will start with cases in base 10 and proceed to arbitrary bases and initial strings.

Finally, in Chapter 4, we will define *good* functions. We will show that good functions are Benford and utilize our result to prove established cases of Benford's Law discussed in [1]. In particular, we will prove that the partition function $p(n)$ and the factorials $n!$ are Benford.

Chapter 2 *Uniform Distribution*

With the introduction of Chapter 1, we are now in place to build the infrastructures for our core theorems. We now introduce some preliminaries on uniform distribution. We will focus on the unit interval $[0,1)$ and start by defining uniform distribution modulo 1. We will then introduce two criteria for determining if sequences are indeed uniformly distributed modulo 1, namely, Weyl's Criterion and Van der Corput's Difference Theorem. We will follow Kuipers and Niederreiter's organization in *Uniform Distribution of Sequences* [5] in presenting this chapter.

2.1 Uniform Distribution Modulo 1

In Chapter 1, we defined the notations for integral and fractional parts of a real number x. We now introduce $\operatorname{frac}(x)$ as a shorthand notation for the fractional part of x. Also, for a real sequence $\{x_n\}$, $N \in \mathbb{N}$, and $E \subseteq \mathbb{R}$, let

$$A\big(E, N, \{x_n\}\big) := \#\left\{1 \leqslant n \leqslant N : x_n \in E\right\}.$$

We remark that $A\big([a,b), N, \{\operatorname{frac}(x_n)\}\big)$ essentially counts the number of elements among the first N entries of $\{x_n\}$ such that their fractional parts are in the interval $[a,b)$.

Proposition 1.1 pertains to the concept that $\{(\log_{10} x_n \bmod 1)\}$ being uniformly distributed over $[0,1)$. We make this idea more precise by defining uniform distribution modulo 1 in Definition 3.

Definition 3. (*Uniform Distribution Modulo 1*) A real sequence $\{x_n\}$ is said to be uniformly distributed modulo 1 if, for all intervals $[a,b) \subseteq [0,1)$,

$$\lim_{N\to\infty} \frac{A\big([a,b), N, \{\mathrm{frac}(x_n)\}\big)}{N} = b - a. \tag{2.1}$$

In other words, this means that the percentage of elements in $\{x_n\}$ with fractional parts in $[a,b)$ is proportionate to the width of $[a,b)$ for every such interval. Now, we define the characteistic function as

$$c_{[a,b)}(x) = \begin{cases} 1, & \text{if } x \in [a,b) \\ 0, & \text{otherwise} \end{cases}$$

so that the expression $\sum_{n=1}^{N} c_{[a,b)}\big(\mathrm{frac}(x_n)\big)$ also evaluates the number of elements among the first N entries of the sequence $\{x_n\}$ such that their fractional part fall in the interval $[a,b)$. Then (2.1) can be rewritten as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} c_{[a,b)}\big(\mathrm{frac}(x_n)\big) = \int_0^1 c_{[a,b)}(x)dx. \tag{2.2}$$

The left hand side equals to that of (2.1), and the right hand side equals $b-a$ by the Fundamental Theorem of Calculus. We will show that (2.2) provides a important tool in proving the following theorem, which shows that the condition on intervals is equivalent to a uniform distribution criterion for all real-valued continuous functions f.

Theorem 2.1. (*Criterion for Uniform Distribution Modulo 1*) The real sequence $\{x_n\}$ is uniformly distributed modulo 1 if and only if for every real-valued continuous function f defined on the closed unit interval $[0,1]$ the following equality holds:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f\big(\mathrm{frac}(x_n)\big) = \int_0^1 f(x)dx. \tag{2.3}$$

Proof. We start proving sufficiency by assuming that $\{x_n\}$ is uniformly distributed modulo 1. Let $P : \{0 = k_0 < k_1 < \cdots < k_m = 1\}$ be a partition of the unit interval $[0,1]$. Consider the step function

$$f_{step}(x) = \sum_{i=0}^{m-1} d_i c_{[k_i,k_{i+1})}(x),$$

where $d_i \in \mathbb{R}$ for all i. Plugging $f_{step}(x)$ into (2.3) is in effect multiplying (2.2) applied to the subinterval $[k_i, k_{i+1})$ by a real constant d_i. Therefore, (2.3) holds for all step functions f_{step} so defined.

Now, let $f : [0,1] \to \mathbb{R}$ be a continuous function. By the definition of Riemann integrals, for any $\epsilon > 0$, there exist 2 step functions f_{step1} and f_{step2} such that $f_{step1} < f < f_{step2}$ on $[0,1]$ and $\int_0^1 (f_{step2}(x) - f_{step1}(x))dx < \epsilon$. Then,

$$\begin{aligned}
\int_0^1 f(x)\,dx - \epsilon \ &\leqslant\ \int_0^1 f_{step1}(x)\,dx = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f_{step1}\big(\mathrm{frac}(x_n)\big) \\
&\leqslant\ \liminf_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f\big(\mathrm{frac}(x_n)\big) \leqslant \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f\big(\mathrm{frac}(x_n)\big) \\
&\leqslant\ \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f_{step2}\big(\mathrm{frac}(x_n)\big) = \int_0^1 f_{step2}(x)\,dx \leqslant \int_0^1 f(x)\,dx + \epsilon.
\end{aligned}$$

Since ϵ can be arbitrarily small, we can produce f_{step1} and f_{step2} accordingly so that $\int_0^1 f_{step1}(x)dx = \int_0^1 f(x)dx = \int_0^1 f_{step2}(x)dx$. Since step functions satisfy (2.3), so does f, and hence the sufficiency.

For necessity, suppose that $\{x_n\}$ is a real sequence satisfying (2.3) for all continuous $f : [0,1] \to \mathbb{R}$. Suppose also that $[a,b) \subseteq [0,1)$. We now show that (2.1) is true for $\{x_n\}$. For arbitrary $\epsilon > 0$, there exist continuous functions g_1, g_2 such that $g_1 \leqslant c_{[a,b)}(x) \leqslant g_2$ on $[0,1]$ and $\int_0^1 \big(g_2(x) - g_1(x)\big)dx \leqslant \epsilon$. Consequently,

$$\begin{aligned} b - a - \epsilon &\leqslant \int_0^1 g_2(x)\ dx - \epsilon \leqslant \int_0^1 g_1(x)\ dx = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g_1\big(\mathrm{frac}(x_n)\big) \\ &\leqslant \liminf_{N\to\infty} \frac{A\big([a,b), N, \{\mathrm{frac}(x_n)\}\big)}{N} \leqslant \limsup_{N\to\infty} \frac{A\big([a,b), N, \{\mathrm{frac}(x_n)\}\big)}{N} \\ &\leqslant \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g_2\big(\mathrm{frac}(x_n)\big) = \int_0^1 g_2(x)\ dx \leqslant \int_0^1 g_1(x)\ dx + \epsilon \leqslant b - a + \epsilon. \end{aligned}$$

Since ϵ can be arbitrarily small, we can produce g_1 and g_2 so that equality holds. Therefore, (2.1) holds for x_n, completing the proof. □

An important corollary of Theorem 2.1 is as follows.

Corollary 2.2. The sequence $\{x_n\}$ is uniformly distributed modulo 1 if and only if for every continuous function $f : \mathbb{R} \to \mathbb{C}$ with period 1, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f\left(x_n\right) = \int_0^1 f(x)dx. \tag{2.4}$$

Proof. We start with sufficiency. Since f is continuous, its real and imaginary parts $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$ must be continuous real-valued functions. Applying Theorem 2.1 on $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$, we conclude that (2.3) holds for f. With the imposed periodicity, it is true that $f(x) = f\big(\mathrm{frac}(x)\big)$, hence (2.4) follows.

For necessity, we assume that (2.4) is true for all continuous complex valued function with period 1 and show that $\{x_n\}$ satisfies (2.1) in Definition 3. This can be done by the same methodology as in the proof of Theorem 2.1. Note that this time we can produce g_1 and g_2 such that $g_1(0) = g_1(1)$ and $g_2(0) = g_2(1)$. In this way, we can extend the domain of g_1 and g_2 to $\mathbb{R}$. Then, following the same proof completes the proof for necessity. □

2.2 Weyl's Criterion

Functions of the form $f(x) = e^{2\pi ikx}$ with integer k are continuous complex-valued functions over $\mathbb{R}$, hence satisfying Corollary 2.2. In other words, if the sequence $\{x_n\}$ is uniformly distributed modulo 1, then (2.4) holds for $f(x)$ so defined. This fact motivates the celebrated Weyl's Criterion by the German mathematician Hermann Weyl. This criterion dramatically simplifies the task of testing whether a sequence is uniformly distributed modulo 1.

Theorem 2.3. (*Weyl's Criterion*) The real sequence $\{x_n\}$ is uniformly distributed modulo 1 if and only if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi ik\cdot x_n} = 0$$

for any nonzero integer k.

Proof. Let $f(x) = e^{2\pi ikx}$. Suppose $\{x_n\}$ is real sequence uniformly distributed modulo

1. By Corollary 2.2, we have that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f\big(x_n\big) = \int_0^1 f(x)dx = \int_0^1 e^{2\pi ikx}dx = 0.$$

For necessity, suppose that $\{x_n\}$ satisfies $\lim\limits_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} e^{2\pi ik\cdot x_n} = 0$. We will show that (2.4) is true for every continuous functions $f : \mathbb{R}\to\mathbb{C}$ with period 1 and then claim that $\{x_n\}$ is uniformly distributed modulo 1 by Corollary 2.2.

We pick an arbitrary $\epsilon > 0$. By Weierstrass Approximation Theorem, there exists a trigonometric polynomial $p(x)$ (a linear combination of terms of the form $e^{2\pi ikx}$ with complex coefficients) such that

$$\sup_{0\leqslant x\leqslant 1} |f(x) - p(x)| \leqslant \epsilon. \tag{2.5}$$

Therefore, by triangular inequality,

$$\begin{aligned}\left|\int_0^1 f(x)dx - \frac{1}{N}\sum_{n=1}^{N} f\big(x_n\big)\right| \quad &\leqslant \quad \underbrace{\left|\int_0^1 \big(f(x) - p(x)\big)dx\right|}_{\text{Term 1}} \\ &\quad + \underbrace{\left|\int_0^1 p(x)dx - \frac{1}{N}\sum_{n=1}^{N} p\big(x_n\big)\right|}_{\text{Term 2}} \\ &\quad + \underbrace{\left|\frac{1}{N}\sum_{n=1}^{N}\big(f\big(x_n\big) - p\big(x_n\big)\big)\right|}_{\text{Term 3}}.\end{aligned}$$

Terms 1 and 3 are less than ϵ because of (2.5). Note that this is true irrelevant of the value of N. Term 2 can be less than ϵ if N is sufficiently large by assumption on $\{x_n\}$. Therefore, $\left|\int_0^1 f(x)dx - \frac{1}{N}\sum_{n=1}^{N} f\big(x_n\big)\right|$ vanishes, completing the proof. □

Weyl's Criterion results in another powerful result, Fejér's Theorem, that will be incorporated in the proof of Section 2.3.

Theorem 2.4. (*Uniform Distribution of $f(n)$*) Let $\{f(n), n = 1, 2, \cdots\}$ be a sequence of real numbers such that $\Delta f(n) = f(n+1) - f(n)$ is monotone as n increases. Let, furthermore,

$$\lim_{n\to\infty} \Delta f(n) = 0 \text{ and } \lim_{n\to\infty} n\,|\Delta f(n)| = \infty.$$

Then the sequence $f(n)$ is uniformly distributed modulo 1.

Proof. We first note that the following inequality holds for any pair of real u and v.

$$\begin{aligned} |e^{2\pi iu} - e^{2\pi iv} - 2\pi i(u-v)e^{2\pi iv}| &= \left|e^{2\pi i(u-v)} - 1 - 2\pi i(u-v)\right| \\ &= 4\pi^2 \left|\int_0^{u-v} (u-v-w)e^{2\pi iw} dw\right| \\ &\leqslant 4\pi^2 \left|\int_0^{u-v} (u-v-w) dw\right| \\ &= 2\pi^2(u-v)^2 \end{aligned} \tag{2.6}$$

Now let $u = kf(n+1)$ and $v = kf(n)$ with $k \in \mathbb{Z}$ being a nonzero constant. Then by dividing (2.6) by $\Delta f(n)$,

$$\left|\frac{e^{2\pi ikf(n+1)}}{\Delta f(n)} - \frac{e^{2\pi ikf(n)}}{\Delta f(n)} - 2\pi ike^{2\pi ikf(n)}\right| \leqslant 2\pi^2k^2\,|\Delta f(n)|$$

where $n \geqslant 1$. Therefore, by triangle inequality, we have

$$\begin{aligned} &\left|\frac{e^{2\pi ikf(n+1)}}{\Delta f(n+1)} - \frac{e^{2\pi ikf(n)}}{\Delta f(n)} - 2\pi ike^{2\pi ikf(n)}\right| \\ &\qquad\qquad \leqslant \left|\frac{1}{\Delta f(n)} - \frac{1}{\Delta f(n+1)}\right| + 2\pi^2k^2\,|\Delta f(n)|, \end{aligned} \tag{2.7}$$

where $n \geqslant 1$. Consequently,

$$\begin{aligned}
&\left|2\pi ik\sum_{n=1}^{N-1} e^{2\pi ikf(n)}\right| \\
&\quad= \left|\sum_{n=1}^{N-1}\left(2\pi ike^{2\pi ikf(n)} - \frac{e^{2\pi ikf(n+1)}}{\Delta f(n+1)} + \frac{e^{2\pi ikf(n)}}{\Delta f(n)}\right) + \frac{e^{2\pi ikf(N)}}{\Delta f(N)} - \frac{e^{2\pi ikf(1)}}{\Delta f(1)}\right| \\
&\quad\leqslant \sum_{n=1}^{N-1}\left|2\pi ike^{2\pi ikf(n)} - \frac{e^{2\pi ikf(n+1)}}{\Delta f(n+1)} + \frac{e^{2\pi ikf(n)}}{\Delta f(n)}\right| + \frac{1}{|\Delta f(N)|} + \frac{1}{|\Delta f(1)|} \\
&\quad\leqslant \sum_{n=1}^{N-1}\left|\frac{1}{\Delta f(n)} - \frac{1}{\Delta f(n+1)}\right| + 2\pi^2k^2\sum_{n=1}^{N-1}|\Delta f(n)| + \frac{1}{|\Delta f(N)|} + \frac{1}{|\Delta f(1)|},
\end{aligned}$$

where the last step uses our result in (2.7). Now, since $\Delta f(n)$ is monotonous, so is $\frac{1}{\Delta f(n)}$. Note that $\sum_{n=1}^{N-1}\left|\frac{1}{\Delta f(n)} - \frac{1}{\Delta f(n+1)}\right|$ is merely the sum of consecutive differences of $\frac{1}{\Delta f(n)}$, it equals to $\left|\frac{1}{\Delta f(N)} - \frac{1}{\Delta f(1)}\right|$ due to monotonicity. By triangular inequality, $\left|\frac{1}{\Delta f(N)} - \frac{1}{\Delta f(1)}\right| \leqslant \frac{1}{|\Delta f(N)|} + \frac{1}{|\Delta f(1)|}$. Thus, dividing the above inequality by $\frac{1}{2\pi|k|N}$ yields

$$\left|\frac{1}{N}\sum_{n=1}^{N-1} e^{2\pi ikf(n)}\right| \leqslant \frac{1}{\pi|k|}\left(\frac{1}{N|\Delta f(1)|} + \frac{1}{N|\Delta f(N)|}\right) + \frac{\pi|k|}{N}\sum_{n=1}^{N-1}|\Delta f(n)|.$$

By assumption, $\lim\limits_{n\to\infty}\Delta f(n) = 0$ and $\lim\limits_{n\to\infty} n\,|\Delta f(n)| = \infty$. The above equation therefore reduces to

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N-1} e^{2\pi ikf(n)} = 0,$$

satisfying Weyl's Criterion. Thus, $\{f(n)\}$ is uniformly distributed modulo 1. □

Refining Theorem 2.4 by replacing $\Delta f(n)$ with $f'(n)$ gives the Fejér's Theorem for differentiable functions. Corollary 2.5 becomes apparent when we take into account the Mean Value Theorem in calculus.

Corollary 2.5. (*Fejér's Theorem*) Let $f(x)$ be a function defined for $x \geqslant 1$ that is differentiable for $x \geqslant x_0$. If $f'(x)$ tends monotonically to 0 as $x \to \infty$ and if

$\lim_{x\to\infty} x\,|f'(x)| = \infty$, then the sequence $f(n), n = 1, 2, \cdots$ is uniformly distributed modulo 1.

Proof. According to the Mean Value Theorem, on any interval $[n, n+1]$, there exists at least one real number $c_n \in (n, n+1)$ such that $f'(c_n) = f(n+1) - f(n) = \Delta f(n)$. By assumption, $\lim\limits_{n\to\infty} \Delta f(n) = \lim\limits_{n\to\infty} f'(c_n) = 0$ and $\lim\limits_{n\to\infty} n\,|\Delta f(n)| = \lim\limits_{n\to\infty} n\,|f'(c_n)| = \infty$. We can now apply Theorem 2.4 to conclude that for $n \geqslant x_0$, $f(n)$ is uniformly distributed modulo 1.

Since there are only finitely many integer n smaller than x_0, they do not affect uniform distribution modulo 1 when n tends to infinity. As a result, the corollary holds for $n = 1, 2, \cdots$. □

We will refer back to Corollary 2.5 in the proof Corollary 2.8 in Section 2.3.

2.3 Difference Theorem

We now proceed to prove Van der Corput's Difference Theorem and an important corollary that will be useful in Chapter 4. Our first step in this section is Lemma 2.6.

Lemma 2.6. (*Van der Corput's Fundamental Inequality*) Let $u_1, \cdots, u_N$ be complex numbers, and let H be an integer with $1 \leqslant H \leqslant N$. Then

$$\begin{aligned} H^2 \left| \sum_{n=1}^{N} u_n \right|^2 \leqslant\ & H(N+H-1) \sum_{n=1}^{N} |u_n|^2 \\ & +2(N+H-1) \sum_{h=1}^{H-1} (H-h) \mathrm{Re} \left(\sum_{n=1}^{N-h} u_n \bar{u}_{n+h} \right) \end{aligned}$$

where $\mathrm{Re}(z)$ is the real part of $z \in \mathbb{C}$, and $\overline{z}$ is the complex conjugate of z.

Proof. Setting $u_n = 0$ for $n \leqslant 0$ and $n > N$, we obtain

$$H\sum_{n=1}^{N} u_n = \sum_{p=1}^{N+H-1}\sum_{h=0}^{H-1} u_{p-h}. \tag{2.8}$$

Squaring both sides of (2.8) and using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
H^2\left|\sum_{n=1}^{N} u_n\right|^2 &\leqslant (N+H-1)\sum_{p=1}^{N+H-1}\left|\sum_{h=0}^{H-1} u_{p-h}\right|^2\\
&= (N+H-1)\sum_{p=1}^{N+H-1}\left(\sum_{r=0}^{H-1} u_{p-r}\right)\left(\sum_{s=0}^{H-1} \overline{u}_{p-s}\right)\\
&= (N+H-1)\sum_{p=1}^{N+H-1}\sum_{h=0}^{H-1} |u_{p-h}|^2\\
&\quad +2(N+H-1)\mathrm{Re}\sum_{p=1}^{N+H-1}\sum_{\substack{r,s=0\\ s<r}}^{H-1} u_{p-r}\overline{u}_{p-s}\\
&= (N+H-1)\left(H\sum_{n=1}^{N}|u_n|^2 + 2\mathrm{Re}\sum\right),
\end{aligned}$$

where $\sum$ in the last row contains terms of the form $u_n\overline{u}_{n+h}$ for $n = 1, 2, \cdots, N$ and $h = r - s = 1, 2, \cdots, H-1$. For fixed n and h within their respective ranges, the pairs of r and s contributing the term $u_n\overline{u}_{n+h}$ are $(h, 0), (h+1, 1), \cdots, (H-1, H-h-1)$. Furthermore, for each of these choices, the value of p is unique. As a result, we have exactly $H - h$ occurrences of $u_n\overline{u}_{n+h}$ in $\sum$. Therefore,

$$\sum = \sum_{h=1}^{H-1}(H-h)\sum_{n=1}^{N} u_n\overline{u}_{n+h}.$$

Note that we can limit n so that $1 \leqslant n \leqslant N-h$ because u_n vanishes for all $n > N$ by our previous definition, hence finishing the proof. □

With the help of Lemma 2.6, we are now in place to prove the difference theorem .

Theorem 2.7. (*Van der Corput's Difference Theorem*) Let $\{x_n\}$ be a given sequence of real numbers. If for every positive integer h the sequence $\{x_{n+h} - x_n, n = 1, 2, \cdots\}$ is uniformly distributed modulo 1, then $\{x_n\}$ is uniformly distributed modulo 1.

Proof. Let $m \neq 0$ be some fixed integer, and let $u_n = e^{2\pi i m \cdot x_n}$ for $n = 1, 2, \cdots$. Dividing the equation in Lemma 2.6 by H^2N^2, one obtains

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i m\cdot x_n}\right|^2 \leqslant \frac{N+H-1}{HN} + 2\sum_{h=1}^{H-1} &\frac{(N+H-1)(H-h)(N-h)}{H^2N^2} \\ &\cdot \left|\frac{1}{N-h}\sum_{n=1}^{N-h} e^{2\pi i m(x_n - x_{n+h})}\right|.\end{aligned} \tag{2.9}$$

By assumption, the sequence $\{x_{n+h} - x_n\}$ is uniformly distributed modulo 1 for any $h \geqslant 1$, so

$$\lim_{N\to\infty} \frac{1}{N-h}\sum_{n=1}^{N-h} e^{2\pi i m(x_n - x_{n+h})} = 0 \tag{2.10}$$

for any $h \geqslant 1$. Substituting the second summand in (2.9) with (2.10),

$$\limsup_{N\to\infty}\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i m\cdot x_n}\right|^2 \leqslant \lim_{N\to\infty}\frac{N+H-1}{HN} = \frac{1}{H}. \tag{2.11}$$

Since $H = 1, 2, \cdots, N$ and N tends to infinity, we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i m\cdot x_n} = 0.$$

By Weyl's Criterion, the sequence x_n is uniformly distributed modulo 1. □

Theorem 2.7 provides yet another criterion for determining whether a sequence is Benford or not. Its consequence, Corollary 2.8 proved below, will be an important tool for proving Theorem 4.1.

Corollary 2.8. Let $k \in \mathbb{N}$ and $f(x)$ be a function defined for $x \geqslant 1$ which is k-times differentiable for all $x \geqslant x_0$ for some $x_0 \in \mathbb{R}^+$. Suppose that $f^{(k)}$ is eventually monotonic. Suppose also that

$$\lim_{x\to\infty} f^{(k)}(x) = 0 \text{ and } \lim_{x\to\infty} x|f^{(k)}(x)| = \infty.$$

Then $\{f(n) : n \in \mathbb{N}\}$ is uniformly distributed modulo 1.

Proof. We will prove the corollary by a simple induction on k. The base case is already taken care of by Fejér's Theorem in Corollary 2.5. For our induction step, we assume the corollary for k, and show that it also holds for f if it is $(k+1)$-times differentiable for $x \geqslant x_0$ and is eventually monotonic with $\lim_{x\to\infty} f^{(k+1)}(x) = 0$ and $\lim_{x\to\infty} x|f^{(k+1)}(x)| = \infty$. For a positive integer h, let $g_h(x) = f(x+h) - f(x)$ for $x \geqslant 1$. Then $g_h^{(k)}(x) = f^{(k)}(x+h) - f^{(k)}(x)$ for $x \geqslant x_0$. By the induction hypothesis, $\{g_h(n), n = 1, 2, \cdots\}$ is uniformly distributed modulo 1. By Van der Corput's result in Theorem 2.7, the corollary holds for $k+1$. □

Chapter 3 *Mathematical Framework for Benford's Law*

In this chapter, we will reformulate Benford's Law with mathematical rigor. We will start by introducing the Benford Space $\mathcal{B}$, and proceed to talk about cases in base 10. Finally, we will generalize Benford's Law to all bases.

3.1 Benford's Law in Base 10

We start by introducing the space of Benford functions.

Theorem 3.1. (*Space of Benford Functions*) Benford functions form a vector space, which we denote by $\mathcal{B}$.

We remark that although the 0 function is in every vector space, it is obviously not Benford. The Benford property only applies to the non-zero functions in $\mathcal{B}$.

We will suppress the complete proof of Theorem 3.1 and only check some of its properties. For instance, $\mathcal{B}$ is closed under scalar multiplication. If $c \in \mathbb{R} - \{0\}$, then $cf(n) \in \mathcal{B}$. This is true because multiplying a nonzero constant c translates the distribution of $\text{frac}\big(\log_k\left(cf(n)\right)\big)$ by $\text{frac}\big(\log_k c\big)$, and it would not change the uniform distribution of the logarithm. Explicitly,

$$\begin{aligned} & \lim_{N\to\infty} \frac{\#\{1 \leqslant n \leqslant N : \operatorname{frac}\big(\log_k \big(cf(n)\big)\big) \in [a,b)\}}{N} \\ = & \lim_{N\to\infty} \frac{\#\{1 \leqslant n \leqslant N : \operatorname{frac}\big(\log_k c + \log_k \big(f(n)\big)\big) \in [a,b)\}}{N} \\ = & \ b-a \end{aligned}$$

As another example, we show that if $f(n)$ is nonzero, then its inverse $1/f(n) \in \mathcal{B}$. This holds since taking reciprocal would leave the uniform distribution unchanged. That is,

$$\begin{aligned} & \lim_{N\to\infty} \frac{\#\{1 \leqslant n \leqslant N : \operatorname{frac}\big(\log_k \big(1/f(n)\big)\big) \in [a,b)\}}{N} \\ = & \lim_{N\to\infty} \frac{\#\{1 \leqslant n \leqslant N : \operatorname{frac}\big(-\log_k \big(f(n)\big)\big) \in [a,b)\}}{N} \\ = & \lim_{N\to\infty} \frac{\#\{1 \leqslant n \leqslant N : \operatorname{frac}\big(\log_k \big(f(n)\big)\big) \in [a,b)\}}{N} \\ = & \ b-a \end{aligned}$$

We will see in later discussions of this chapter that for all functions of $\mathcal{B}$, Benford's Law must be true for any initial string of digits in any base [1].

Recall that in scientific notation, we write all real numbers in the format $x = S(x) \cdot 10^k$, where $S(x) \in [1, 10)$. To quickly get our hands on the leading digit of x, we introduce the logarithm mapping $\varphi : \mathbb{R} \to [0, 1)$ defined as

$$x \mapsto \log_{10} x \bmod 1.$$

Under logarithm mapping, $\varphi(x) = \log_{10}\big(S(x) \cdot 10^k\big) \bmod 1 = \big(\log_{10} S(x) + k\big) \bmod 1$. Since k is always an integer, $\varphi(x) = \log_{10} S(x) \bmod 1$, the mantissa of x. As $S(x)$ goes from 1 to 10, $\varphi(x)$ varies from 0 to 1. Therefore, the position of $\varphi(x)$ in $[0, 1)$ is sufficient for determining the leading digit of x. For instance, if x has leading digit 1, then $1 \leqslant S(x) < 2$ and $0 = \log_{10} 1 \leqslant \varphi(x) < \log_{10} 2$. Similarly, if x has leading digit

D_1, then $D_1 \leqslant S(x) < D_1 + 1$ and $\log_{10} D_1 \leqslant \varphi(x) < \log_{10}(D_1 + 1)$. Likewise, we find the leading digit of x based on where $\varphi(x)$ lands on the interval $[0, 1)$, as in Figure 3.

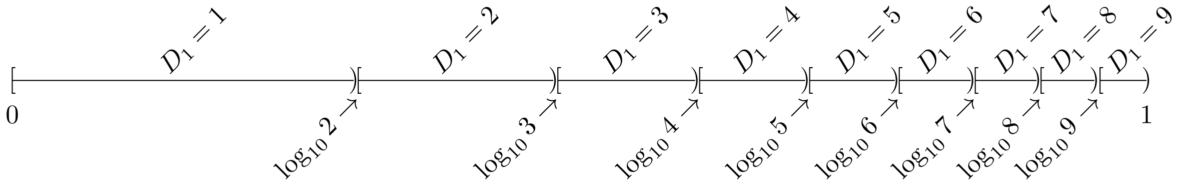


Figure 3. *Logarithm Map*

In the light of uniform distribution modulo 1, we now give a revised version of Proposition 1.1 credited to Persi Diaconis.

Proposition 3.2. (*Diaconis*) A sequence $a(n) \in \mathcal{B}$ if and only if $\log_{10}\big(a(n)\big)$ is uniformly distributed modulo 1.

A generalized version of Proposition 3.2 will be proven in Section 3.2, and we shall delay the proof until then.

3.2 Generalization to All Bases

Recall that the base of a number system is the number of distinct digits in its representation. Thus far, our discussion has been limited to base 10 only, where the distinct digits are 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. In Section 1.2.3, we considered the scale-invariant explanation of Benford functions. Our discussion in Section 1.2 Benford sequences should be scale invariant. Particularly, if the sequence $\{x_n\}$ is Benford under base 10, it should be Benford under any bases. This conjecture motivates us to generalize Benford's Law to arbitrary bases in Definition 4.

Definition 4. (*Generalized Benford's Law*) For a sequence of positive integers $\{x_n\}$, let

$$B(d, N, k; \{x_n\}) = \frac{\#\{n \leqslant N : \text{first digits of } x_n \text{ in base } k \text{ are the string } d\}}{N}.$$

If

$$\lim_{N\to\infty} B(d, N, k; \{x_n\}) \equiv \big(\log_k(d+1) - \log_k d\big) \pmod 1$$

for all $k \geqslant 2$, then we say that $\{x_n\} \in \mathcal{B}$.

To illustrate this, we consider the binary system. By Definition 4, the percentage of numbers starting with the string $d = 101$ is $\log_2(101_2 + 1) - \log_2(101_2) = \log_2(6) - \log_2(5) = 26.3\%$, and the percentage of numbers starting with the string $d = 10101$ is $\log_2(10101_2 + 1) - \log_2(10101_2) = \log_2(22) - \log_2(21) = 6.7\%$. The same is true no matter what base we choose.

We now state and prove Diaconis' Criterion for Benford functions in Theorem 3.3. Note that its special case in base 10 has already served as a fundamental theorem in our discussion so far.

Theorem 3.3. (*Diaconis' Criterion for Benford*) Suppose $\{x_n\}$ is a real sequence, then $\{x_n\} \in \mathcal{B}$ if and only if $\{\log_k x_n\}$ is uniformly distributed modulo 1 for all k.

Proof. In base $k \in \mathbb{Z}$, we can write any real number x as $S(x) \cdot k^n$, with $S(x)$ being the generalized significand and n the exponent of x. In the light of Section 3.1, taking the logarithm of x, we discover that the first digits of x_n are the string d if and only if $\log_k S(x) \bmod 1 \in \big[\log_k d, \log_k(d+1)\big)$. Therefore, we seek to prove that $\{\log_k x_n\}$ is uniformly distributed modulo 1 if and only if the significands of $\{x_n\}$ are Benford.

Since the probability of $\log_k x_n \in [a,b) \subset [0,1)$ is simply that of $\log(x_n) \in [0,b)$ minus that of $\log(x_n) \in [0,a)$, without loss of generality, we only need to show that the probability of $\log(x_n) \in [0,t)$ is t for any $t \in [0,1)$. Let $s \in [1,10)$ be any significand. By previous discussion, we have that

$$\{n \leqslant N : \log_k x_n \bmod 1 \in \big[0, \log_k s\big)\} = \{n \leqslant N : S(x_n) < s\}.$$

It then follows that

$$\lim_{N\to\infty} \frac{\#\{n \leqslant N : \log_k x_n \bmod 1 \in \big[0, \log_k s\big)\}}{N} = \lim_{N\to\infty} \frac{\#\{n \leqslant N : S(x_n) < s\}}{N} \tag{3.1}$$

If $\{\log_k x_n\}$ is uniformly distributed modulo 1, then the left hand side of (3.1) implies that

$$\lim_{N\to\infty} \frac{B(d,N,k;\{x_n\})}{N} = \log_k(d+1) - \log_k d,$$

namely x_n is Benford. Conversely, if $\{x_n\} \in \mathcal{B}$, then the right hand side equals $\log_k s$. This means that the probability of $\log_k x_n \in [0, \log_k s)$ is $\log_k s$, implying the uniform distribution of $\{\log_k x_n\}$. □

3.3 Asymptotic Property

Lemma 3.4 is an asumptotic property of Benford functions. We shall see later that it will be useful in the proof of Theorem 4.1 in Chapter 4.

Lemma 3.4. (*Asymptotically Benford*) If $f(n) \in \mathcal{B}$ and $f(n) \sim g(n)$, then $g(n) \in \mathcal{B}$.

Proof. Suppose that $f(n) \in \mathcal{B}$, then $\{\log_k \big(f(n)\big)\}$ is uniformly distributed modulo 1 by Theorem 3.3. By Definition 3,

$$\lim_{N\to\infty} \frac{\#\{1 \leqslant n \leqslant N : \mathrm{frac}\big(\log_k \big(f(n)\big)\big) \in [a,b)\}}{N} = b - a.$$

It follows from Theorem 2.3 that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k \cdot f(n)} = 0$$

for any nonzero $k \in \mathbb{Z}$. Since $f(n) \sim g(n)$, for any $\epsilon > 0$, we can produce N_0 such that $|g(n) - f(n)| < \epsilon$. Then, we have that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k \cdot g(n)} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N_0 - 1} e^{2\pi i k \cdot g(n)} + \lim_{N\to\infty} \frac{1}{N} \sum_{n=N_0}^{N} e^{2\pi i k \cdot g(n)}.$$

Since the first sum on the right hand side is bounded, the limit vanishes. The second term reduces to

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=N_0}^{N} e^{2\pi i k \cdot g(n)} \leqslant e^{\pm\epsilon} \lim_{N\to\infty} \frac{1}{N} \sum_{n=N_0}^{N} e^{2\pi i k \cdot f(n)}.$$

With $|\epsilon|$ being arbitrarily small, this term goes to 0 by assumption on f. We have therefore proved that $g(n) \in \mathcal{B}$ by Weyl's Criterion. □

Chapter 4 *Benford Arithmetic Functions*

This final chapter exhibits arithmetic functions that conform to Benford's Law. We will define good functions, and then use our results for uniform distribution modulo 1 to prove that good functions are Benford. This is important because it enables us to find a new class of Benford functions, among which are the partition function $p(n)$ and the factorials $n!$.

4.1 Good Functions

We now introduce another criterion for Benford based on the idea of good functions defined in Definition 5.

Definition 5. (*Good Functions*) An integer-valued function $a(n)$ is *good* whenever $a(n) \sim b(n)e^{c(n)}$ and the following conditions are satisfied.

1. There exists some integer $h \geqslant 1$ such that $c(n)$ is h-differentiable and $c^{(h)}(n)$ tends to zero monotonically for sufficiently large n.
2. $\lim\limits_{n\to\infty} n|c^{(h)}(n)| = \infty$.
3. $\lim\limits_{n\to\infty} \frac{D^{(h)}\log b(n)}{c^{(h)}(n)} = 0$, where $D^{(h)}$ denotes the h^{th} derivative.

The following theorem proven by Anderson, Rolen and Stoehr [1] would serve as a convenient tool in proving a number of arithmetic functions in the Benford space $\mathcal{B}$. It is important because it immediately proves that the partition function $p(n)$ and the factorials $n!$ are Benford, as we shall see in Sections 4.2 and 4.3.

Theorem 4.1. (*Good Implies Benford*) If $\{x_n\}$ is good, then $\{x_n\} \in \mathcal{B}$.

Proof. Since $\{x_n\}$ is good, we can find functions $b(n)$ and $c(n)$ that satisfy the conditions in Definition 5. Then, Lemma 3.4 suggests that we only have to show that $b(n)e^{c(n)}$ is Benford. By Theorem 3.3, we only need to show that $\log b(n) + c(n)$ is uniformly distributed modulo 1.

By conditions 1 and 2 in Definition 5, $c(n)$ satisfies the limit assumptions of Theorem 2.8, and is therefore uniformly distributed modulo 1. By condition 3 in Definition 5, $\lim_{n\to\infty} \frac{D^{(h)} \log b(n)}{c^{(h)}(n)} = 0$. Therefore adding $\log b(n)$ would leave the uniform distribution of $c(n)$ unaffected, hence completing the proof. □

4.2 The Partition Function $p(n)$ is Benford

The partition function $p(n)$ for non-negative integer n counts the number of ways to write n as the sum of a non-increasing sequences. For instance, when $n = 4$, there are 5 distinct ways to decompose n as desired. That is,

$$\begin{aligned} 4 &= 4 \\ &= 3 + 1 \end{aligned}$$

$$
\begin{aligned}
&= \quad 2+2 \\
&= \quad 2+1+1 \\
&= \quad 1+1+1+1.
\end{aligned}
$$

Therefore, $p(4) = 5$. Similarly, we can compute that $p(5) = 7$, $p(6) = 11$ and so on. It is worth noting that as n increases, $p(n)$ grows exponentially. When $n = 100$, $p(n) = 190569292$, and when $n = 200$, $p(n) = 3972999029388$.

We now list the first digit distribution of $p(n)$ below. Tables 4 and 5 are reproductions of Tables 1 and 2 in [1], with Table 4 for the case of base 10 and Table 5 for base 2. In both cases, the data suggests that $p(n) \in \mathcal{B}$. In the light of Theorem 4.1, the following corollary automatically holds due to the Hardy-Ramanujan asymptotic partition formula.

x	$d = 1$	2	3	4	5	6	7	8	9
10^2	0.330	0.160	0.140	0.090	0.070	0.060	0.070	0.050	0.030
10^3	0.305	0.177	0.127	0.094	0.076	0.068	0.057	0.052	0.044
10^4	0.302	0.177	0.126	0096	0.078	0.067	0.057	0.051	0.046
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞?	0.301?	0.176?	0.125?	0.097?	0.079?	0.067?	0.057?	0.051?	0.046?

Table 4. $B\big(d, x, 10; p(n)\big)$

x	$d = 100$	101	110	111
200	0.285	0.270	0.205	0.225
400	0.308	0.273	0.209	0.205
600	0.313	0.267	0.217	0.198
800	0.314	0.263	0.219	0.201
1000	0.315	0.262	0.220	0.200
5000	0.321	0.264	0.222	0.194
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞?	0.322?	0.263?	0.222?	0.192?

Table 5. $B\big(d, x, 2; p(n)\big)$

Corollary 4.2. The partition function $p(n) \in \mathcal{B}$.

Proof. By Hardy-Ramanujan in [4], we know that

$$p(n) \sim \frac{1}{4n\sqrt{3}} \cdot e^{\pi\sqrt{2n/3}}.$$

By Definition 5, $p(n)$ is good, and by Theorem 4.1, $p(n) \in \mathcal{B}$. □

4.3 Factorials are Benford

We have exhibited the first digit distributions in both Table 2 and Figure 2 (d). In fact, with the help of Stirling's formula, we can easily prove that $n! \in \mathcal{B}$.

Corollary 4.3. The sequence of factorials $n! \in \mathcal{B}$.

Proof. Stirling's formula in [9] states that

$$n! \sim \sqrt{2\pi} \cdot n^{n+1/2} e^{-n}.$$

By Definition 5, $n!$ is good, and by Theorem 4.1, $n! \in \mathcal{B}$. □

Bibliography

[1] T. Anderson, L. Rolen, and R. Stoehr, *Benford's Law For Coefficients of Modular Forms and Partition Functions*, Proceedings of the American Mathematical Society, 139 (2011), pp. 1533–1541.

[2] F. Benford, *The Law of Anomalous Numbers*, Proceedings of the American Philosophical Society, 78 (1938), pp. 551–572.

[3] W. Feller, *An Introduction to Probability Theory and Its Appliations*, vol. 2, John Wiley & Sons, New York City, New York, 2 ed., 1971.

[4] G. H. Hardy, *Ramanujan: Twelve Lectures on Subjects Suggested by His Life and Work*, AMS Chelsea Publishing, New York City, New York, 3 ed., 1999.

[5] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Dover Publications, Inc., Mineola, New York, 2006.

[6] R. J. Larsen and M. L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, Prentice Hall, Boston, Massachusetts, 5 ed., 2012.

[7] S. J. Miller, ed., *Benford's Law: Theory and Applications*, Princeton University Press, Princeton, New Jersey, 2015.

[8] S. Newcomb, *Note on the Frequency of Use of the Different Digits in Natural Numbers*, American Journal of Mathematics, 4 (1881), p. 3940.

[9] E. M. Stein and R. Shakarchi, *Complex Analysis*, Princeton University Press, Princeton, New Jersey, 2003.